[ICON 2007: 5th International Conference on Natural Language Processing, Hyderabad, Januray 2007]

Semantically Relatable Sequences in the Context of Interlingua Based Machine Translation

Rajat Kumar Mohanty
Department of Computer Science & Engineering
Indian Institute of Technology Bombay
Mumbai-400076 INDIA
Email: rkm@cse.iitb.ac.in

M. Krishna Prasad
Department of Computer Science & Engineering
Indian Institute of Technology Bombay
Mumbai-400076 INDIA
Email: mkrishna@cse.iitb.ac.in

Lakshmi Narayanaswamy
Department of Computer Science & Engineering
Indian Institute of Technology Bombay
Mumbai – 400076 INDIA
Email: nlakshmi@cse.iitb.ac.in

Pushpak Bhattacharyya
Department of Computer Science & Engineering
Indian Institute of Technology Bombay
Mumbai-400076 INDIA
Email: pb@cse.iitb.ac.in

Abstract

In Interlingua based machine translation source language sentences have to be converted to a semantic representation- often a *meaning graph* with concept nodes and semantic relations- from which the target language sentences are produced. We argue that towards the meaning graph generation, a necessary step is to detect the sentence constituents which participate in semantic linkages. Semantic linkages are of the form, *relation(entity1, entity2).* Before creating the semantic linkages it is necessary to detect *(entity1, entity2)* which we call a *Semantically Relatable Sequence (SRS)*. SRS computation makes use of NLP tools like the parser and NLP resources like the WordNet and OALD. Once SRSs are generated, we have covered a considerable distance to the translation.

For evaluating the efficacy of the SRS generation system, we show that the system accurately produces the shallow semantic role labels of 92,310 sentences of the FrameNet corpus. It is emphasized that the system ultimately is designed to produce deep semantic role labels in the framework of Universal Networking Language (UNL) which is a recently proposed interlingua. An important by-product of the work is the fact that the costly resource of semantically role labeled corpus can be obtained at least partially *automatically* through our system.

1 Introduction

In the context of interlingua based Machine Translation, there is a fundamental question to be addressed:

> *Given a sentence containing the words*
> *(W_1-W_2-W_3…-W_n)*
> *what are the semantically relatable sequences?*

Our work is motivated by seeking an answer to this question. Language Phenomena like *movement, empty pronominals, copular constructs, partitives and small clauses* are some of the challenges to be negotiated on the way. Ascertaining that traditional phrase structured and dependency parsers cannot meet these challenges, we employ basic level probabilistic parsing [Charniak, 2000], subcategorization properties of content words and the lexical resources like WordNet [Miller, 2003] and OALD [Hornby, 2001] to identify *semantically relatable sequences (SRSs)* which can ultimately be treated to generate semantic role labels within the UNL framework [Uchida *et. al.,* 1999].

Given an input sentence, the system breaks the constituents into one of the three basic semantically relatable sequence frames such as **<entity1 entity2>** or **<entity1 functional-element entity2>** or **<functional-element entity>**, where the entities can be single words or more complex sentence parts (such as embedded clauses). Ultimately, these sequences are to be labeled with either abstract semantic roles (like *agent* (agt), *object* (obj), *goal*(gol), *instrument* (ins), *etc*.), or are expressed in terms of attributes such as *@topic*, *@present*, *@past*, *@proximate*, *@interrogative*, *@passive*, *etc*. Figure 1 shows the abstract process of interlingua- based machine translation where SRS and UNL are two intermediate stages.

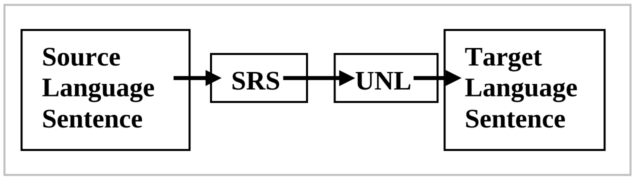

Figure 1: Abstract Process of Interlingua-based MT

For experimentation and evaluation, we have used the FrameNet Corpus [Baker *et. al.*, 1998] which comes semantically annotated in terms of the arguments and adjuncts of the main verb. We create the **Gold Standard SRSs** out of this corpus taking verbs, nouns and adjectives as targets. The total number of sentences is **92,310.** The system's accuracy of SRS identification currently is **82%**.

In what follows, we give in section 2 a brief description of the Universal Networking Language (UNL) framework. Semantically Relatable Sequences (SRS) are defined in section 3. Some non-trivial language phenomena along with the corresponding SRSs are presented in section 4. The process of evolving the computational strategy from linguistic insights are described in section 5 and the implementation in section 6. Section 7 presents the evaluation and experimental results. Section 8 is on related work. Section 9 concludes the paper.

2 Universal Networking Language

UNL is an electronic language for computers to express and exchange information [Uchida *et. al.*, 1999]. It consists of *universal words (UW), relations, attributes* and the *UNL knowledge base (KB)*. The UWs constitute the vocabulary of UNL, relations and attributes the syntax, and the UNLKB the semantics of the framework. UNL represents information sentence by sentence as a hyper-graph[1] with concepts as nodes and relations as arcs. The root node is the main verb of the sentence. Figure 2 represents the UNL graph for sentence (1).

(1) The professor is teaching NLP to the CS students.

In the figure, the arcs are labeled with the relation labels *agt (agent), gol (final state), mod (modifier)* and *obj (object)*. The nodes *professor(icl>person), teach(icl>instruct), NLP(icl>discipline), CS(icl>department)* and *student(icl>person)* are the Universal Words (UW), *i.e.,* disambiguated concepts. The restrictions in parentheses serve to make the sense unique. The lexical relation *icl* stands for *included-in. @entry (the root node), @present (tense), @progress (aspect)* and *@def (definite)* are attributes which provide further information about how the concept is being used in the specific sentence.

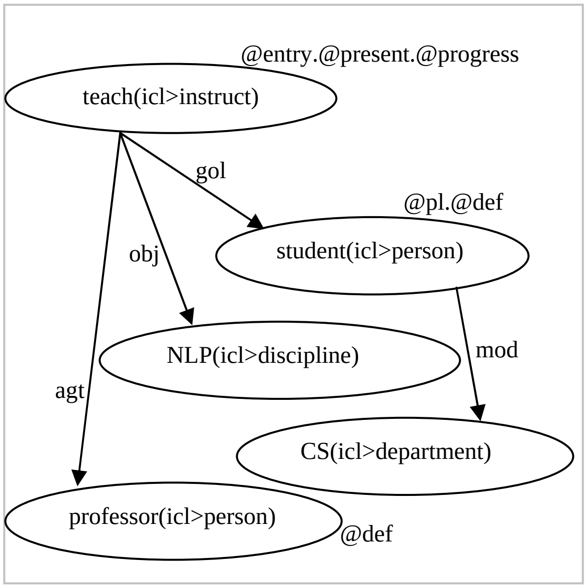

Figure 2: UNL Graph for the Sentence (1)

3 Semantically Relatable Sequences

Words in natural languages are broadly classified into two categories: content words and function words. Sentence (1) above contains five content words- *professor, teaching, NLP, CS*, and *students* and four function words- *the, is, to* and *the*. The UNL graph in figure 2 suggests that the content and function words can be regarded as constituting the following sequences:

(2) *a.* *(professor, teaching)*
b. *(teaching, NLP)*
c. *(teaching, to, students)*
d. *(CS, students)*
e. *(is, teaching)*
f. *(the, professor)*
g. *(the, students)*

We postulate that a sentence needs to be broken into sequences of at most three forms, as shown in (3).

[1] The nodes themselves can be graphs- hence the name *hyper-graph..*

(3) a. (CW, CW)
b. (CW, FW, CW)
c. (CW, CW)

The notation *FW* refers to all function words in a language. *CW* refers to either a simple content word or a compound concept which is expressed by a notation called *scope node* and denoted as *SCOPExx* where xx are 2 digits. We hereby develop the notion of semantically relatable sequences postulated by Mohanty *et.al.* (2005).

Definition: *A semantically relatable sequence (SRS) of a sentence is a group of words in the sentence, not necessarily consecutive, that appear in the semantic graph of the sentence as linked nodes or nodes with speech act labels*.

We contend that once the SRSs have been produced, we have covered a considerable distance towards translation. The generation of a Hindi sentence from English is illustrated *via* SRS and UNL in (4-7).

(4) John eats rice with a spoon.

(5) [SRS]
(John, eats.@entry)
(eats, rice)
(eats, with, spoon)
(a, spoon)
[\SRS]

(6) [UNL]
agt(eat.@entry.@present, John)
obj(eat.@entry.@present, rice)
ins(eat.@entry.@present, spoon.@indef)
[\UNL]

(7) जॉन चम्मच से चावल खाता है।
John spoon with rice eat BE-PRES
“John eats rice with a spoon.”

4 SRS and Language Phenomena

Many complex language phenomena have to be handled on the way to SRS generation, some of which are described now to give a glimpse of the inherent complexity involved. Space constraint does not permit an exhaustive description. In what follows, the illustrative sentences are numbered *(N)a.*- where *N* is a digit- and the corresponding SRSs are numbered as *b*.

4.1 SRS and Movement Phenomena

Topicalization: In (8a), the object NP is topicalized.

(8) a. The problem, we solved.
b. (we , solved.@entry)------------(CW, CW)
(solved.@entry , problem)-----(CW,CW)
(the, problem)-------------------(CW,CW)

Preposition Stranding: A stranded preposition is one which has been separated from its complement by movement operation. The complement of the preposition is topicalized. For a sentence like (9a), the SRSs are given in (9b).

(9) a. John, we laughed at.
b. (we , laughed.@entry)---------(CW, CW)
(laughed.@entry,at, John)---(CW, FW, CW)

Relative Clauses: In (10a), a defining-relative clause is illustrated. The relative pronoun refers to the grammatical object of the relative clause which, in fact, involves movement. The fact is discussed in some detail in section 5.2.

(10) a. John told a joke which we had already heard.
b. (John, told.@entry) -------------------(CW, CW)
(told.@entry, :01) ---------------------(CW,CW)
SCOPE01(we,had,heard.@entry)-------(CW, FW,CW)
SCOPE01(already,heard.@entry)-------(CW,CW)
SCOPE01(heard@entry,which,joke)----(CW,FW,CW)
SCOPE01(a, joke)-------------------------(FW,CW)

Note the use of the notation *SCOPE01* to represent the complete clause *we had already heard.*

Interrogatives: An interrogative always involves a movement operation.

(11) a. Who did you refer her to?
b. (did , refer.@entry.@interrogative)-------(FW,CW)
(you, refer.@entry.@interrogative)--------(CW,CW)
(refer.@entry.@interrogative , her)--------(CW,CW)
(refer.@entry.@interrogative,to,who)---(CW,FW,CW)

4.2 SRS and Empty Pronominals

These are sentences in which there are elements not visible in the usual text.

To-infinitivals: Theoretically speaking, a to-infinitival clause has an empty pronominal, called PRO, which is covertly present as the grammatical subject of the clause. Detection of the PRO elements in a to-infinitival clause, and subsequent resolution of the co-indexing of the PRO element are not trivial in SRS generation.

(12) Bill was wise to sell the piano.

In (12), the PRO element is co-indexed with *Bill* in the *underlying structure*. The SRSs generated for the sentence in (12) are given in (13).

(13) (wise.@entry , SCOPE01)---------------(CW,CW)
SCOPE01(sell.@entry , piano)---------(CW,CW)
(Bill, was, wise.@entry)--------------(CW, FW,CW)

SCOPE01(Bill, to, sell.@entry)-----(CW, FW,CW)
SCOPE01(the, piano) ----------------(FW,CW)

In (13), the entire to-infinitival clause appears under a SCOPE and this is referred to in the SRS *(wise.@entry, SCOPE01)*. The entity *Bill* acts as the grammatical subject of the matrix clause as well as the entity participating in the SRS: *SCOPE01(Bill, to, sell.@entry)* by virtue of its co-indexing with the PRO element.

Gerundial Constructs: *Verbal gerunds* refer to an action, whereas *nominal gerunds* refer to a fact. Structurally, there are a number of properties that show that verbal gerunds have the syntax of verbs whereas nominal gerunds have the syntax of basic nouns. One of the variants of the verbal gerunds shows the presence of an empty pronominal PRO. (14a) illustrates a verbal gerund in which the empty pronominal PRO is subject-controlled.

(14) a. The cat leapt down spotting a thrush on the lawn.
b. (The, cat) --------------------------------(FW, CW)
(cat, leapt.@entry) -----------------------(CW, CW)
(leapt.@entry , down) -------------------(CW, CW)
(leapt.@entry , SCOPE01)--------------(CW, CW)
SCOPE01(cat, spotting.@entry)--------(CW,CW)
SCOPE01(spotting.@entry,thrush)-----(CW,CW)
SCOPE01(spotting.@entry,on,lawn)---(CW,FW,CW)
SCOPE01(a, thrush) -------------------------(FW, CW)
SCOPE01(the, lawn) ------------------------(FW, CW)

4.3 SRS and PP-Attachment

We focus our attention on the particular frame [V-NP1–P-NP2]. The prepositional phrase attachment sites under various conditions and associated heuristics are enumerated in [Mohanty *et. al.*, 2005]. Using these heuristics, the sentence in (15a) can be broken into SRSs as shown in (15b).

(15) a. John cracked the glass with a stone.
b. (John, cracked.@entry)--------------(CW,CW)
(cracked.@entry, glass)-------------(CW,CW)
(cracked.@entry, with, stone)----(CW,FW,CW)
(a, stone)------------------------------(FW,CW)
(the,glass)----------------------------(FW,CW)

This was a sample of the language phenomena addressed in our system. We now show how the linguistic insights so obtained are used to obtain actual algorithms.

5 From Linguistics to Computation

A probabilistic parser [Charniak, 2004], the English WordNet 2.0 [Miller, 2003] and Oxford Advanced Learners Dictionary [Hornby, 2001] are used to generate the semantically relatable sequences.

Parse trees generated from Charniak Parser bring out the basic structure of the sentence and the tag information at every node helps identify the relatable sequences. Movement/copying operation is applied in a reverse way (say, the *rightward movement*) to project the underlying positions of content words in a sentence.

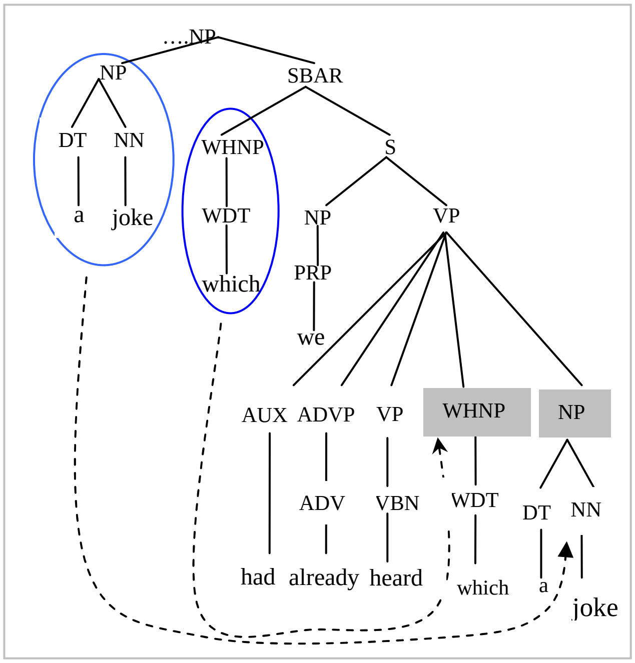

Figure 3: Partial Tree showing Rightward Movement

The movement operation is implemented as follows. The discussion uses the sentence (11a) *John told a joke which we had already heard.*

a. The clause boundary is identified as the NP above the SBAR that denotes the relative clause.
b. If there exists an NP between the relative pronoun and the verb of the relative clause, the pronoun acts as the object of the relative clause and hence must be moved. In this case, the NP *we* (acting as the subject) helps decide that a movement operation is necessary.
c. The NP that is the antecedent of the relative pronoun is determined as the closest NP sibling of the SBAR clause. Here, *joke* is determined as the antecedent.

d. The relative pronoun and its NP antecedent is added to the end of the verb phrase in the relative clause.

The movement of the antecedent NP is not motivated by the transformational theory, but for the purpose of SRS generation. The resulting augmented tree is given in figure (3).

5.1 Creation of a visible node for an invisible element

The parse tree of the gerundial sentence (14a) is augmented, as shown in Figure 4.

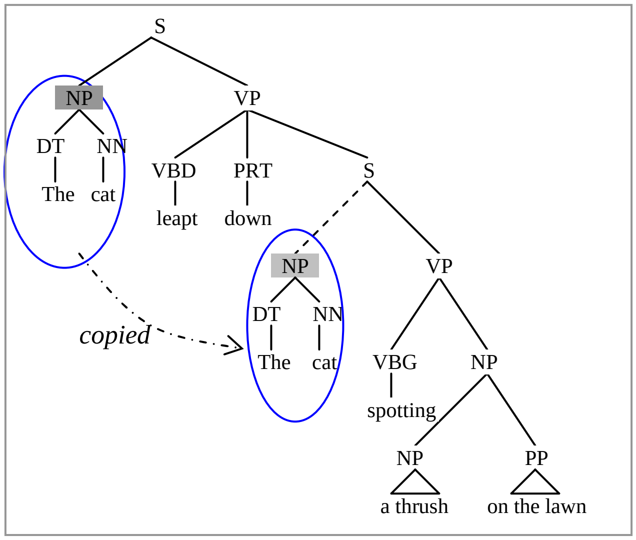

Figure 4: Parse Tree for a Verbal Gerundial Clause

The VBG node under the VP has no AUX sister indicating that the VP heads a verbal gerundial clause. As discussed earlier, a verbal gerundial clause contains an empty pronominal PRO and it can be subject controlled, object controlled or have an arbitrary PRO (*for instance*, it is subject controlled in 14a).

Hence, the following modifications are done to the parse tree as shown in Figure 4:

a. The clause boundary is the S node, which is labeled with the head SCOPE to indicate that it is a compound concept.

b. A VP node is checked if it contains a sibling TO node with VB sister to indicate that it is a to-infinitival clause or it contains a VBG node with no AUX sister to indicate that it is a case of verbal gerund.

c. The duplication and insertion of an NP node with head *cat* (depicted by shaded nodes in Figure 4) as a sibling of the VP node with the head VBG *spotting* is done to bring out the existence of a semantic relation between *cat* and *spotting*.

6 Implementation

A high-level overview of the SRS Generator system is presented in figure 5.

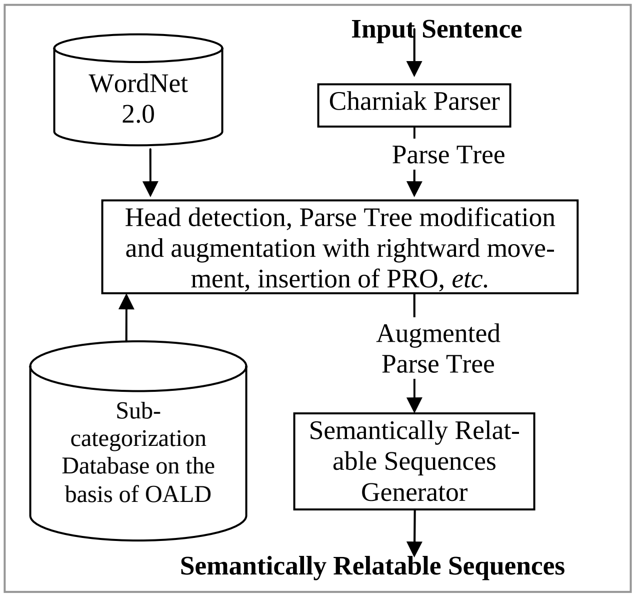
Input Sentence
WordNet 2.0
Charniak Parser
Parse Tree
Head detection, Parse Tree modification and augmentation with rightward movement, insertion of PRO, etc.
Augmented Parse Tree
Sub-categorization Database on the basis of OALD
Semantically Relatable Sequences Generator
Semantically Relatable Sequences

Figure 5: An overview of the System Architecture

After parsing the sentence, the *head determination* module uses a bottom-up strategy to determine the head word for every node in the parse tree. This module processes the children of every node using heuristics that rely on tagsets of potential heads of that node. There are some special cases, such as:

a. For SBAR node, the scope handler module is invoked to decide the kind of clause, scope creation points and heads of nodes in its locality.

b. A VP node is checked for the following:

i. For to-infinitival clauses as well as gerundial clauses, PRO insertion is done depending on whether it can be co-indexed with some element in the parse tree, or is missing completely.

ii. If the copula *be* is the head of VP and is followed by an adjectival or noun phrase, the head of the latter phrase is taken to be the head of the predicate, *e.g., She is famous.*

c. NP nodes are checked for of-PP cases and conjunctions under them, which lead to scope creation.

d. SBARQ nodes are also handled, causing movement of the wh-word to the appropriate verb phrase.

The above strategy gives rise to a case by case algorithm which is again omitted due to space constraint.

7 Evaluation

7.1 Creation of Testdata

For our experiment, we use the FrameNet corpus [Baker *et. al.,* 1998], a semantically annotated corpus, as the testdata. 92310 sentences are taken for evaluation. The *gold standard SRSs* are created automatically from the FrameNet corpus taking verbs, nouns and adjectives as the targets. The statistics of data are in Table 1.

It should be kept in mind that the FrameNet role labeled corpus provides the top level entities which are arguments of the main verb. Our system generates much finer level entities. Thus for the sentence *The judge dismissed the {lawsuit filed in the Yorkshire County Court}* the fragment corpus shows the entire unit in *{}* as the argument without breaking it down into finer constituents. But our system details out this complex constructs. The evaluation designed by us checks to see that at least the FrameNet level argument generation succeeds on a large corpus.

7.2 Design of the Evaluation Function

Looking upon SRSs as sets of *items to be matched* we use *a Weighted Intersection-Union Similarity Measure* to evaluate our system. The matching of the main verb and the generation of the actual attribute *@entry* are very important and are given a weightage of 25% each. Let the set of entities connected to the main verb be denoted as E_G and E_P for the *gold standard SRSs* and the *produced SRSs* respectively. We employ the expression

$$E = \frac{|E_G \bigcap E_P|}{|E_G|}$$

Then the score is computed as

*Score= 0.25*V + 0.25*N + 0.5*E*

Where *V=1 if main verbs match and N=1 if @entry is produced.*

7.3 Experiments and Top Level Statistics

Table 1 gives the target wise (*i.e. verbs, nouns, adjectives*) score. Table 2 gives the same for the verbs of high frequency in the *gold standard*. Figure 6 shows scores of ten high frequency verb groups [Levin, 1993] and Figure 7 shows the same for noun groups [Miller, 2003]. The overall accuracy is 82%**.**

Target Types	Total # sentences	Score
Verbs as targets	37,984	0.766
Nouns as targets	37,240	0.867
Adjectives as target	17,086	0.840

Table 1: Top Level Statistics

Verbs	Frequency In FrameNet	Avg. #Gold SRS	Max. #Gold SRS	Score
swim	280	2.3	5	0.709
depend	215	1.3	3	0.804
look	187	2.3	4	0.835
roll	173	1.9	4	0.7
rush	172	2.3	4	0.775
phone	162	2.5	4	0.695
reproduce	159	2.3	4	0.797
step	159	2.5	4	0.795
urge	157	2.5	4	0.765
avoid	152	2.4	4	0.789

Table 2: Score for High Frequency Verbs

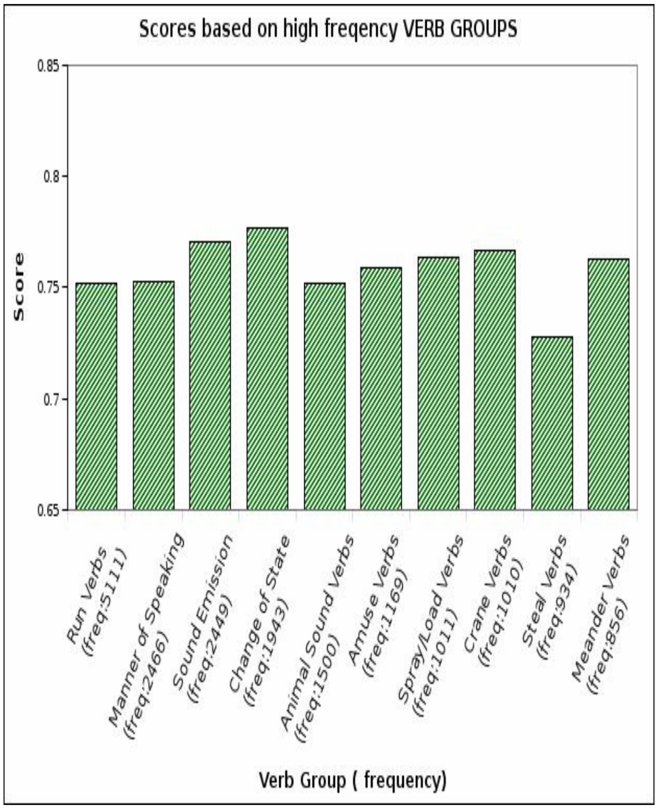

Figure 6: Scores of 10 Verb Groups of High Frequency in the Gold Standard

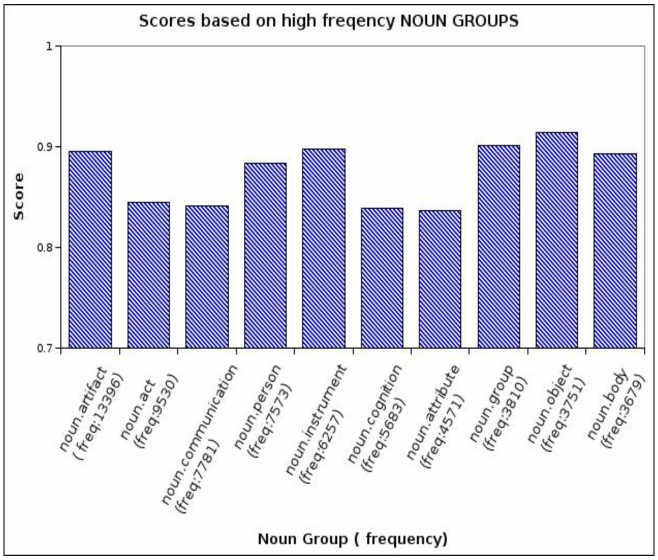

Figure 7: Scores of 10 Noun Groups of High Frequency in the Gold Standard

We have done the error analysis to some extent, and discovered that accuracy reduction is caused by:

(i) Some of the sentences in FrameNet are wrongly labeled. For instance, in the sentences "*It would seem that some 38% of the parishes surveyed by the Commission depend upon the goodwill of such people, and there are fewer of them in rural than in urban areas*.", *depend* is the target verb for the noun *Commission*. However, for the noun *Commission* the target verb is *surveyed*, which is accurately generated by our system.

(ii) Charniak parser fails to handle sentences with length greater than or equal to 200 words.

(iii) The parser has inherent limitations. For instance, in the sentence "*She stood up and began pacing restlessly to and fro along the balcony*", the parse tree assigns an NN tag to *pacing* instead of a VBG tag. Not only that, it also assigns VP to the parent of NN.

(iv) The Parser is unable to handle some of the language phenomena like *gapping*.

8 Related Work

Interlingua representations have been widely studied in the machine translation literature [Hutchins and Somers, 1992]. One of early noteworthy interlingua based MT systems is Atlas-II [Uchida, 1989]; the comparison of the interlingua approach to the more widespread transfer approach is done in Boitet [1988]; the consequence of language divergence on interlingua has been recently studied in Dave *et. al.* [2002].

As far as shallow semantic parsing is concerned, Semantic Role Labeling (SRL) has been studied by several researchers, such as, Baker *et. al*. [1998], Bejan *et. al*. [2004], Gildea and Jurafsky [2002], Mohanty *et. al*. [2005]. Hacioglu (2004) shows SRL on the basis of dependency trees.

9 Conclusion and Future work

We have described here our work on obtaining *semantically relatable sequences (SRS)* which are vital intermediaries towards semantic graph generation. The semantic graphs in our case are UNL graphs. The high accuracy of results on a set of highly representative language phenomena, validate the stand that it pays to form a good understanding of linguistic constructs and translate these insights into computational systems.

Many interesting applications are possible from this work. We have tried SRS based search in IR and have observed the high precision it consistently attains compared to an ordinary search engine like *Lucene*. However, the recall is poor. But the *Mean Average Precision Scores (MAP)* and *R-prec* scores are again consistently higher.

The obvious future work is to obtain the actual semantic roles from SRSs. This would employ a knowledge base with the properties of relation labels. The use of SRSs in directly creating approximate translations in a target language is being investigated. Automatic learning of SRSs is another interesting problem.

References

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet Project. Proceedings of COLING.

C. A. Bejan, A. Moschitti, P. Morarescu, G. Nicolae, and S. Harabagiu. 2004. Semantic Parsing Based on FrameNet. Proceedings of SENSEVAL- 3.

Christian Boitet. 1988. Pros and Cons of the Pivot and Transfer Approaches in Multilingual Machine Translation. In Klaus Schubert and Toon Witkam (eds.), *Recent Developments in Machine Translation.* Dan Maxwell, Foris, Dordrecht.

Eugene Charniak. 2004. Statistical Parser. http://www.cs.brown.edu/~ec/#software

Shachi Dave, Jignashu Parikh and Pushpak Bhattacharyya. 2002 Interlingua Based English Hindi Machine Translation and Language Divergence, *Journal of Machine Translation (**JMT**)*, 17.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. http://www.cs.rochester.edu/~gildea/gildea-cl02.pdf

W. John Hutchins and Harold L. 1992. Somers. *An Introduction to Machine Translation.* Academic Press, London.

Kadri Hacioglu. 2004. Semantic Role Labeling using dependency trees. http://sds.colorado.edu/SERF/papers/ hacioglu_coling2004.pdf

A. S. Hornby. 2001. *Oxford Advanced Learners' Dictionary of Current English.* OUP.

Levin, Beth. 1993. *English verb Classes and Alternation*. The University of Chicago Press, Chicago.

George Miller. 2003. *WordNet 2.0*. http://wordnet.princeton.edu

Rajat Kumar Mohanty, Anupama Dutta and Pushpak Bhattacharyya. 2005. Semantically Relatable Sets: Building Blocks for Knowledge Representation. Proceeding of 10th MT Summit, Phuket, Thailand (12-15 September 2005)

Hiroshi Uchida. 1989. ATLAS-II: A machine translation system using conceptual structure as an interlingua. In *Proceedings of the Second Machine Translation Summit*, Tokyo.

Hiroshi Uchida, M. Zhu, and T. Della. Senta. 1999. UNL: A Gift for a Millennium. The United Nations University, Tokyo. http://www.undl.org/ publications/gm/top.htm